Intermediate R – Data Wrangling

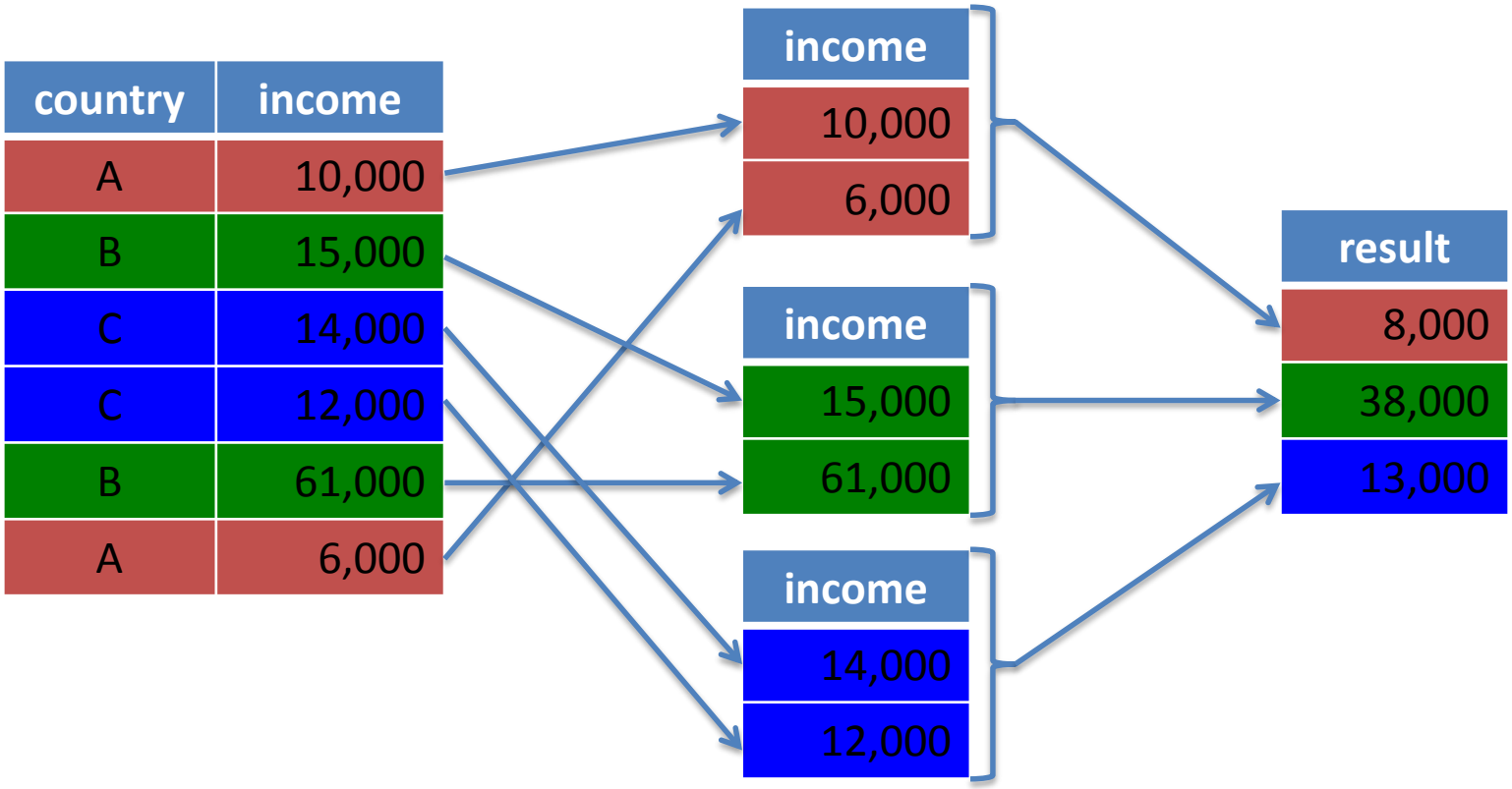
Clark Pixton

January 8, 2014

15.S60

tapply() – summarizing by group

- "What is the mean income of each country?"
- tapply(income, country, mean)



Assignment 1

- What is the average departure delay (DepDelayMinutes) by weekday (DayOfWeek)?
- What is the average taxi-in time by airport (using the “Dest” field)?
- Extra: What is the proportion of cancelled flights by airline (using the “Cancelled” field)? Which airlines have the highest and lowest proportions of cancelled flights?
  - Hint 1: Use == instead of = to check equality
  - Hint 2: The average of TRUE/FALSE values is the proportion that are TRUE

Assignment 2

- One simple way to measure the “skew level” of a distribution is subtract the median from the mean. Write a function that calculates this measure of skew for arrival delays (ArrDelayMinutes) and use tapply to calculate it for each carrier.
  - Hint: use the “median” function.
- Extra: What is the most common Origin-Destination pair for each carrier?
  - Hint: use the paste() function. What would you give as the first argument for tapply?

Split-Apply-Combine

- spl = split(flightsDelayInfo, Origin)

Origin	Carrier Delay	ArrDelay Minutes	…
BOS	0	15	…
ATL	1	12	…
ATL	44	44	…
ATL	13	27	…
BOS	33	50	…
BOS	24	27	…
BOS	0	10	…
ATL	0	5	…
BOS	0	7	…
BOS	17	17	…

Origin	Carrier Delay	ArrDelay Minutes	…
ATL	1	12	…
ATL	44	44	…
ATL	13	27	…
ATL	0	5	…

Origin	Carrier Delay	ArrDelay Minutes	…
BOS	0	15	…
BOS	33	50	…
BOS	24	27	…
BOS	0	10	…
BOS	0	7	…
BOS	17	17	…

Split-**Apply**-Combine

- spl2 = lapply(spl, delay.prop.df)

Origin	Carrier Delay	ArrDelay Minutes	…
ATL	1	12	…
ATL	44	44	…
ATL	13	27	…
ATL	0	5	…

Origin	prop.carrier	..
ATL	0.659	..

Origin	Carrier Delay	ArrDelay Minutes	…
BOS	0	15	…
BOS	33	50	…
BOS	24	27	…
BOS	0	10	…
BOS	0	7	…
BOS	17	17	…

Origin	prop.carrier	..
BOS	0.587	..

Split-Apply-**Combine**

- flights.delay.info = rbind(spl2[[1]], spl2[[2]], …)
- flights.delay.info = do.call(rbind, spl2)

Origin	prop.carrier	..
ATL	0.659	..

Origin	prop.carrier	..
BOS	0.587	..

Origin	prop.carrier	..
ATL	0.659	..
BOS	0.587	..

Assignment 3

- From the flightsFlown data frame, create a data frame called carrier.info, where each row corresponds to one carrier (airline). Include the following variables in your new data frame:
  - carrier: The carrier code
  - mean.arr.delay: Average arrival delay time (using ArrDelayMinutes)
  - longest.delay: Longest flight delay for the month
  - most.common.origin: most common origin for the carrier